Berry Project Report

Zijie Huang

2020/10/20

1. Abstract

This project seeks to explore the **Berry** dataset which were collected from the USDA database selector: https://quickstats.nass.usda.gov. Data were first cleaned and reorganized. Then, data were explored and visualized.

2. Data Cleaning

2.1 Procedure

The unique value of each column were checked. The berries data had only 8 out of 21 columns containing meaningful data. In this project, only strawberry data were considered. Aftering checking each meaningful column, **Year**, **State**, **Value** columns were found to be well-organized. However, column **Data Item**, **Domain**, and **Domain Category** are messy. I chose to split them by every special symbol such as **"-"** and **","**. Then, I extracted all meaningful value without overlap and classified them into new columns. Finally, new column **Production**, **Avg**, **Measures**, **Materials** and **Chemicals** were built. All missing values were substituted by one unit blank space.

2.2 Cleaned Data Brief

Production contains the type of information such as production and yield. **Avg** indicates whether the value is calculated on average. **Measures** contains the unit of **Value**. **Materials** contains the specific information of the matters used. **Chemicals** contains the types of materials such as fungicide and herbicide.

Year	State	production	Avg	Measures	Materials	Chemicals	Value
2019	CALIFORNIA	ACRES HARVESTED			NOT SPECIFIED	TOTAL	35400
2019	CALIFORNIA	ACRES PLANTED			NOT SPECIFIED	TOTAL	36000
2019	CALIFORNIA	PRODUCTION		MEASURED IN $	NOT SPECIFIED	TOTAL	2221320000
2019	CALIFORNIA	PRODUCTION		MEASURED IN CWT	NOT SPECIFIED	TOTAL	20500000
2019	CALIFORNIA	YIELD		MEASURED IN CWT / ACRE	NOT SPECIFIED	TOTAL	580

3. Exploratory Data Analysis

By looking at the **Production**, data can be divided into two parts. One part is about the general data such as the amount of harvested strawberry or planted strawberry. The other part is about the chemicals that were applied to strawberry. Then, we first look at the first part.

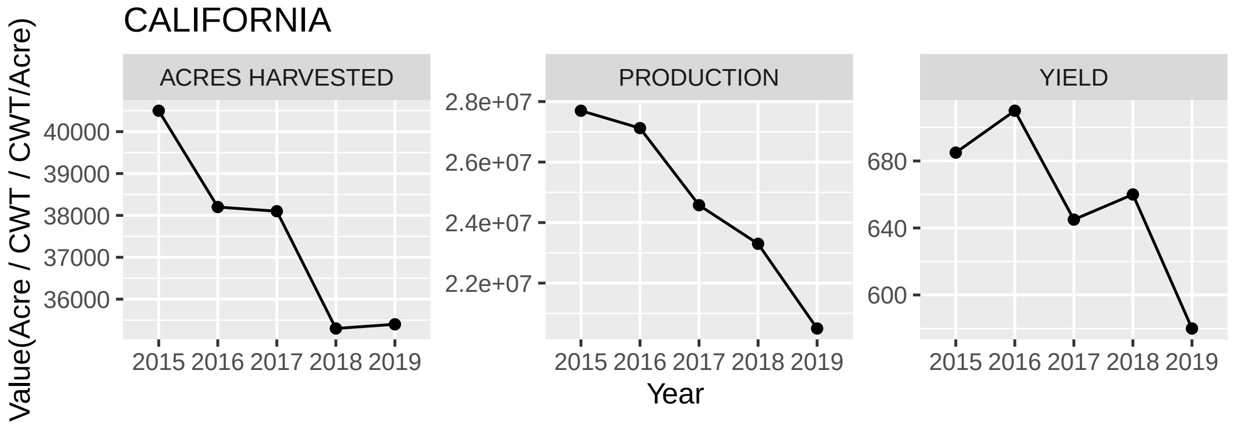
3.1 General Data

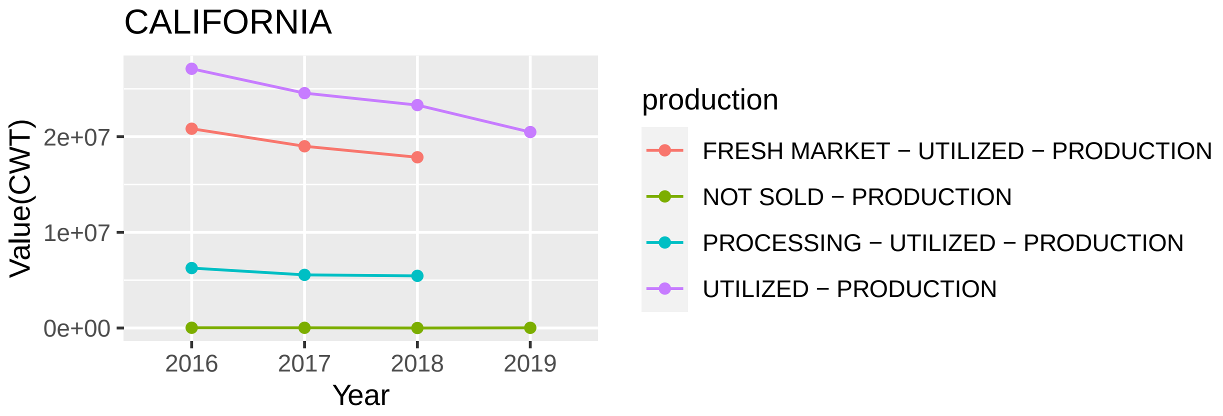
In this part, there are 10 different types of value as shown below. By observation, the relationship between them were found.
1. "Production" = "Harvested" × "Yield".

2. “Utilized-Production” = “Fresh Market-Utilized-Production” + “Processing-Utilized-Production”. Therefore, plots for these data were considered. These plots for each state and year can be found in the shiny application. Below is a simple demonstration.

```
## Types of value:
```

```
##  [1] "ACRES HARVESTED"
##  [2] "ACRES PLANTED"
##  [3] "PRODUCTION"
##  [4] "YIELD"
##  [5] "FRESH MARKET - PRODUCTION"
##  [6] "FRESH MARKET - UTILIZED - PRODUCTION"
##  [7] "PROCESSING - UTILIZED - PRODUCTION"
##  [8] "NOT SOLD - PRODUCTION"
##  [9] "PROCESSING - PRODUCTION"
## [10] "UTILIZED - PRODUCTION"
```





3.2 Chemical Data

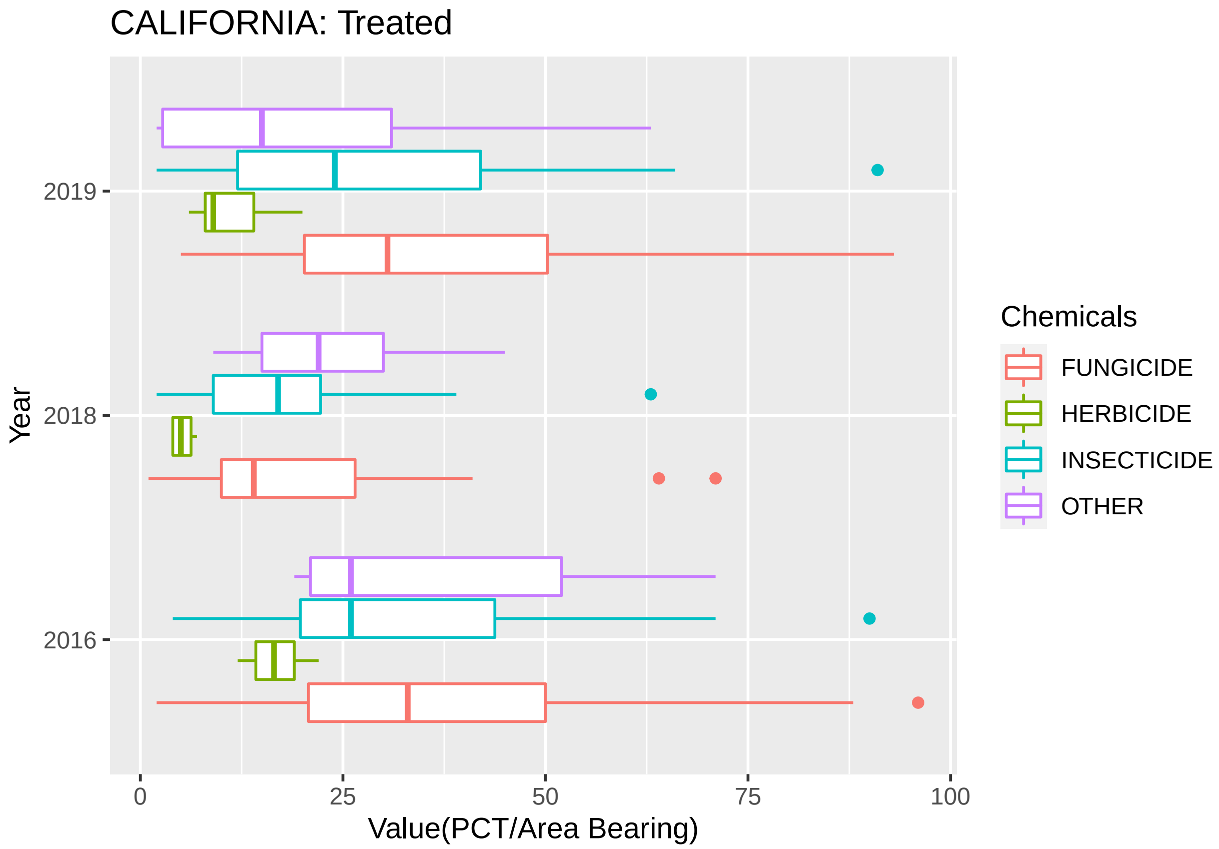
In this part, there are two types of value as shown below. For both applications strawberry and treated strawberry, there are eight types of chemicals as shown below. Since there are many data for each types of chemicals and putting different types of chemicals in the same plot is good for comparison, boxplot of chemicals in each year were considered. These plots can be found in the shiny application. Below is a simple demonstration.

```
## Types of value:
```

```
## [1] "BEARING - APPLICATIONS" "BEARING - TREATED"
```

```
## Types of chemicals:
```

```
## [1] "FUNGICIDE"              "HERBICIDE"              "INSECTICIDE"
## [4] "OTHER"                  "(NITROGEN) FERTILIZER"  "(PHOSPHATE) FERTILIZER"
## [7] "(POTASH) FERTILIZER"    "(SULFUR) FERTILIZER"
```



4. Summary

4.1

With cleaned data, we explored the data from two aspects. One is general data and the other one is chemical data.
From the perspective of intergrity, many data were missing especially for some states, i.e. Michigan, Ohio and Pennsylvania. On the contrary, data for California and Florida are very complete. Apparently, there exists emphasis on some states. For chemical part, there are very few data about fertilizer which makes it hard to study it. There are only four states contain these data also with emphasis on California and Florida.
From the perspective of data plot, there are distinguishable changes over years for many variables. For the boxplot of chemicals, there exists differences between different types of chemicals.

4.2 Difficult Points

During this project, I encountered a few difficult points.
1. The format of column **Value** in the csv file is automatically set as numerical value. Therefore, when trying to convert column **Value** from character to double, errors occurred. I solved this by changing the

format of **Value** from numerical value to general format in the csv file.
2. In the shiny application, how to set dynamic plots which all the plots depends on the input value is the second difficut point. I solved this by using **uiOutput**, **taglist** and **map**. Build a function to render plot and use taglist and map to use this function inside the uiOutput.